Connections and symbols II

AT1

Emmanuel Dupoux

Cogmaster, 2015
summary of preceding session

• how could a physical system ‘think’?
  – reasoning (mathematical, common sense)
  – performing complex cognitive tasks

• philosophical issue: “computational reduction”
  – reduction of unboundedly complex behavior to the combination of simple ones
    • simple set of primitive processes
    • finite set of data types
    • a finite set of operations that combine the primitive processes to make more complex ones

• Scientific issue: precise (formal, quantitative) theories of cognition
  – what computational mechanisms underly complex behaviors (like language, reasoning, etc)?
    • Symbolic IA: (sequential & deterministic) computations with symbols and rules
      – eg: Turing machines, rewrite rules, finite state automata
    • Connexionist IA: (parallel & stochastic) computation with (continuous valued) neurone-like units
      – eg: Multilevel Perceptrons, Boltzman machines

• Technological issue: usable systems (IA)
McClelland & Rumelhart’s (1986) Parallel Distributed Processing

• Cognition involves the spreading of activation, relaxation, statistical correlation.
• Represents a method for how symbolic systems might be implemented
  – Hypothesized that apparently symbolic processing is an emergent property of subsymbolic operations.
• Advantages
  – Fault tolerance & graceful degradation
  – Can be used to model learning
  – More naturally capture nonlinear relationships
  – Fuzzy information retrieval
  – bridges the gap with real neural processing
just the root form of a work similar to those studied by Kohonen (1). The model is designed to capture the phonological structure of the root form of a word and the past-tense verb form. The model consists of two basic parts: (a) a simple pattern associator that learns the relationships between the base form and the past-tense form, and (b) a modifiable connections network that generates as its output. The response strengths for the high-frequency regular verbs are not much different between these alternatives. In most cases the response strengths for the past +ed alternative grows rapidly to over 0.80 while the strengths of the regularized alternatives jump up.

The response strength for the correct response. After about Trial 30, the response strength for the correct alternative drops rapidly to less than 1. The strength of response for several different response alternatives as a function of trials. The response strengths to correspond roughly to relative response probabilities.

FIGURE 1. The basic structure of the model.

FIGURE 2. The network of modifiable connections.

FIGURE 3. The response strengths for the high-frequency regular verbs as a function of trials.

FIGURE 4. The percentage of correct features for regular and irregular high-frequency verbs.

FIGURE 5. The response strengths for the high-frequency regular verbs as a function of trials.

Rumelhart & McClelland (1986)
The critique: Pinker & Mehler (1988)

- Lachter & Bever: connectionist theories are a return to associationism (Chomsky vs Skinner revisited)
- Pinker & Prince: connectionist models of the capacity to derive the past tense of English verbs is inadequate
  - rules: wug \rightarrow wugged
  - exceptions: put \rightarrow put, go\rightarrow went, dig\rightarrow dug
- Fodor & Pylyshyn: connectionist theories are inadequate models of language and thought
Position of the problem: classical theories vs connectionism

Agree:
- both classical theories & connectionism are representationists (they assign some ‘meaning’ to the elements – symbols or nodes)

Disagree
- classical theory encode structural relationships and processes (eg, constituents, variables, rules)
- connectionnists only encode causal relationships and processes (x causes y to fire)

Arguments against connectionnist systems: mental representation and processes are structure sensitive

- combinatorial semantics
  - semantics of « J. loves M. » derived from semantics of « J. », « loves » and « M. »
- productivity
  - the list of thoughts/sentences is not finite (I can construct new thoughts with old ones)
- systematicity
  - I construct them in a systematic way
  - eg: « x loves M. » (where x can be any proper noun)
  - eg: If I can think « J. loves M. », I can think « M. loves J. »
- recursivity & constituent structure:
  - If I can think « P. thinks that M. is nice » I can think « J thinks that P thinks that M is nice »

-> connectionist systems have none of the above properties
Fodor & Pylyshyn (cont)

• Objections to symbolic/classical systems
  – rapidity of cognitive processes/neural speed
  – difficulty of pattern recognition/content based retrieval in conventional architectures
  – committed to rule vs exception dichotomy
  – inadequate for intuitive/nonverbal behavior
  – acutely sensitive to damage/noise (vs graceful degradation)
  – storage in classical systems is passive
  – inadequate account of gradual/frequency based application of rules
  – inadequate account of nondeterminism
  – no account of neuroscience
  → none of these arguments are valid or relevant

• CONCLUSIONS
  1. current connectionist theories are inadequate
  2. if they were to be made adequate they would be mere implementation of classical architecture
in brief

• The Fodor & Pylyshyn challenge:
  – (current) connectionist architectures fail to capture complex behaviors
  – (future) connectionist architectures are ‘mere’ implementation of symbolic architectures
• **Structure of Smolensky**
  – representing structures by fillers and roles
    • examples: trees, lists, etc
  – tensor products and filler/role binding (definition)
    • local, semilocal and distributed
  – unbinding (exact and self addressed)
  – capacity and graceful saturation
  – continuous and infinite structures
  – binding and unbinding networks
  – analogy between binding units and hebb weights
  – example of a stack
  – structured roles
the Smolensky response: tensor products

Paul loves Mary -> loves(Paul, Mary)
pred = loves, arg1 = Paul, arg2 = Mary
pred*loves + arg1*Paul + arg2*Mary

Paul loves Mary
Mary is loved by Paul

Mary loves Paul
Paul is loved by Mary

→ Problem of tensor product representations: exponential with sentence complexity

• extensions:
  – implementation of a phonological theory (Optimality Theory) in a tensor product network with energy relaxation
    • see the Harmonic Mind (Smolensky & Legendre, 2012)
  – Escaping the explosion in nb of neurons: holographic reduced representations
    • define $A \ast B$ as an operation that preserves the dimensions (eg xor, circular convolution)

Plate (1995)
Elman

- structure of the paper
  - representing time
  - SRN architecture
  - xor through time
  - badiiguuuu
  - word segmentation (15 words)
  - part of speech (13 categories, 29 words, 15 sentence templates)
Figure 2. A simple recurrent network in which octivotions are copied from hidden layer to context layer on a one-for-one basis, with fixed weight of 1.0. Dotted lines represent trainable connections.

Input, and also the previous internal state of some desired output. Because the patterns on the hidden units are saved as context, the hidden units must accomplish this mapping and at the same time develop representations which are useful encodings of the temporal properties of the sequential input. Thus, the internal representations that develop are sensitive to temporal context; the effect of time is implicit in these internal states. Note, however, that these representations of temporal context need not be literal. They represent a memory which is highly task- and stimulus-dependent.

Figure 7. Hierarchical cluster diagram of hidden unit activation vectors in simple sentence prediction task. Labels indicate the inputs which produced the hidden unit vectors: inputs were presented in context, and the hidden unit vectors averaged across multiple contexts.

Several points should be emphasized. First, the category structure appears to be hierarchical. Thus, "dragons" are large animals, but also members of the class [human, +animate] nouns. The hierarchical interpretation is achieved through the way in which the spatial relations (of the representations) are organized. Representations that are near one another in the representational space form classes, while higher level categories correspond to larger and more general regions of this space. Second, it is also true that the hierarchy is "soft" and implicit. While some categories may be qualitatively distinct (i.e., very far from each other
• extensions of Elman’s SRN
  – computational capacity of SRN

  – reservoir computing
    • http://reservoir-computing.org
Conclusions

• What about the F&P Challenge?
  – tensor products are an interesting implementation/alternative to symbolic systems
  – recurrent networks could also be an alternative, but much less understood

• The hidden debate
  – innate vs learner structures (to be continued…)

Conclusions

- empirical impact of the debate
  - past tense in English
    - rule: play->played, fax->faxed
    - exceptions: sing->sang, put->put
    - Pinker & Prince (1988)
    - procedural vs declarative memory (Ullman et al, 1997; Pinker & Ullman, 2002)

Conclusions

• empirical impact of the debate (cont)
  – statistical learning vs algebraic learning in infants
      Peña, M., Bonatti, L., Nespor, M., Mehler J. (2002). Science
  – exemplar-based versus abstract representations
    • object recognition (Biederman & Gerhardstein, 1993),
      face recognition, speech recognition (eg Goldinger, 1988; Johnson 1997, Pierrehumbert 2001)

Pierrehumbert, J. (2001). In J. Bybee and P. Hopper (eds.), Frequency and the Emergence of Linguistic Structure
Current trends

• very deep hierarchical networks
  – trained in a heavily supervised fashion
  – but, seem to correlate with human performance and monkey IT neurons


1000 categories
100.000 test
1M training
Google net: 6.67% error
Humans: 5%

Google net, Szegedy et al, 2014
<table>
<thead>
<tr>
<th>robin</th>
<th>cheetah</th>
<th>armadillo</th>
<th>lesser panda</th>
</tr>
</thead>
<tbody>
<tr>
<td>centipede</td>
<td>peacock</td>
<td>jackfruit</td>
<td>bubble</td>
</tr>
<tr>
<td>king penguin</td>
<td>starfish</td>
<td>baseball</td>
<td>electric guitar</td>
</tr>
<tr>
<td>freight car</td>
<td>remote control</td>
<td>peacock</td>
<td>African grey</td>
</tr>
</tbody>
</table>
Following up on recurrent networks

• **LSTM**
  (long short term memory)


• replaces HMM for language models

  Graves, Mohamed, Hinton (2013). Speech Recognition with Deep Recurrent Neural Networks. *ICASSP.*

• **text generation**

Vinyals et al (2014)
see also Donahue et al (2014)
• playing with RNNs
    (the unreasonable effectiveness of RNNs)

• memory networks
  – Weston Chopra & Bordes (2015)
  – question answering

PANDARUS:
Alas, I think he shall be come approached and the day
When little strain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
• attention networks
  – Xu, Ba, .. Bengio 2015

Figure 3. Examples of attending to the correct object (white indicates the attended regions, underlines indicated the corresponding word)

A woman is throwing a frisbee in a park.  A dog is standing on a hardwood floor.  A stop sign is on a road with a mountain in the background.

A large white bird standing in a forest.  A woman holding a clock in her hand.  A man wearing a hat and a hat on a skateboard.
Neural Turing Machine
– learns copy, sort

Evolutions of the symbolic systems

• probabilistic /bayesian frameworks
→ symbolic systems that learn

Le jeu des nombres

- input: nombre entre 1 et 100
- output: “oui” ou “non”

- Tâche d’induction:
  - Observer quelques exemples de ‘oui’
  - Juger si des nouveaux nombres auraient pu être “oui” ou “non”.

Le jeu des nombres

Exemples de "oui"  
Jugements de généralisation  
\((N = 20)\)

60

Similarité diffuse
Le jeu des nombres

<table>
<thead>
<tr>
<th>Exemples de &quot;oui&quot;</th>
<th>Jugements de généralisation ((N = 20))</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>Similarité diffuse</td>
</tr>
<tr>
<td>60 80 10 30</td>
<td>Règle: &quot;multiples de 10&quot;</td>
</tr>
</tbody>
</table>
Le jeu des nombres

Exemples de "oui"
(N = 20)

60

60 80 10 30

60 52 57 55

Jugements de généralisation

Similarité diffuse

Règle: "multiples de 10"

Similarité focalisée: nombres proches de 50-60
Modèle Bayesien

- $H$: Espace d’hypothèse des concepts possibles
  - $h_1 = \{2, 4, 6, 8, 10, 12, \ldots, 96, 98, 100\}$ ("nombres pairs")
  - $h_2 = \{10, 20, 30, 40, \ldots, 90, 100\}$ ("multiples de 10")
  - $h_3 = \{2, 4, 8, 16, 32, 64\}$ ("puissances de 2")
  - $h_4 = \{50, 51, 52, \ldots, 59, 60\}$ ("nombres entre 50 et 60")
  - ...
Modèle Bayesien

- \( H \): Espace d’hypothèse des concepts possibles
  - \( h_1 = \{2, 4, 6, 8, 10, 12, \ldots, 96, 98, 100\} \) (“nombres pairs”)
  - \( h_2 = \{10, 20, 30, 40, \ldots, 90, 100\} \) (“multiples de 10”)
  - \( h_3 = \{2, 4, 8, 16, 32, 64\} \) (“puissances de 2”)
  - \( h_4 = \{50, 51, 52, \ldots, 59, 60\} \) (“nombres entre 50 et 60”)
  - \ldots

- \( X = \{x_1, \ldots, x_n\} \): \( n \) exemples du concept \( C \).

- Évaluer l’hypothèse \( h \) étant donné \( X \):

\[
p(h \mid X) = \frac{p(X \mid h)p(h)}{\sum_{h' \in H} p(X \mid h')p(h')}
\]

- \( p(h) \) [“prior”]: biais à priori
- \( p(X \mid h) \) [“likelihood”]: information statistique dans les exemples
- \( p(h \mid X) \) [“posterior”]: degré de confiance que \( h \) est l’extension de \( C \).
## Comparaison modèle-humain

<table>
<thead>
<tr>
<th>Exemples de “oui”</th>
<th>Jugements ( (N = 20) )</th>
<th>Modèle Bayesien ( (r = 0.96) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td><img src="#" alt="Graph" /></td>
<td><img src="#" alt="Graph" /></td>
</tr>
<tr>
<td>60 80 10 30</td>
<td><img src="#" alt="Graph" /></td>
<td><img src="#" alt="Graph" /></td>
</tr>
<tr>
<td>60 52 57 55</td>
<td><img src="#" alt="Graph" /></td>
<td><img src="#" alt="Graph" /></td>
</tr>
</tbody>
</table>
Comparaison modèle-humain

<table>
<thead>
<tr>
<th>Exemples de “oui”</th>
<th>Jugements ((N = 20))</th>
<th>Modèle Bayesien ((r = 0.93))</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td><img src="image1" alt="Histogram" /></td>
<td><img src="image2" alt="Histogram" /></td>
</tr>
<tr>
<td>16 8 2 64</td>
<td><img src="image3" alt="Histogram" /></td>
<td><img src="image4" alt="Histogram" /></td>
</tr>
<tr>
<td>16 23 19 20</td>
<td><img src="image5" alt="Histogram" /></td>
<td><img src="image6" alt="Histogram" /></td>
</tr>
</tbody>
</table>

from Josh Tennenbaum (MIT)
• Modèles linguistiques probabilistes/Natural Language processing
  • Automates finis stochastiques
  • Grammaires Context Free Probabilistes


• Learning of a syntactic/semantic parser
  • from pairs of sentence-meaning candidates


• from pairs of questions and answers (plus a database of facts)

Not covered here

• other connectionnist architectures
  – Kohonen’s maps (competitive learning) (Kohonen, 1982)
  – Adaptive Resonance Theory (Grossberg, 1976)
  – Reinforcement learning (Barto, Sutton, Anderson, 1983)

• other computational frameworks
  – more Probabilistic/Bayesian frameworks
  – Predictive Coding/Free Energy (Friston 2009)